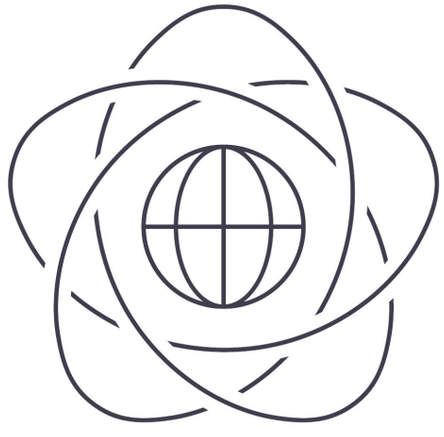


Case Studies 2022L

Variable importance

Apr 14, 2022

Questions?

- Which variables are the most important?
- What to pay attention to?
- When modeling the price of a property, we would like to know what so much impact on the price, whether it is the area or maybe the year of construction?
- When modeling the credit risk, we consider what influenced the fact that customers do not get a loan.

Usage purposes of measure

- **Model simplification:** variables that do not influence a model's predictions may be excluded from the model.

- **Model exploration:** comparison of variables' importance in different models may help in discovering interrelations between the variables. Also, the ordering of variables in the function of their importance is helpful in deciding in which order should we perform further model exploration.

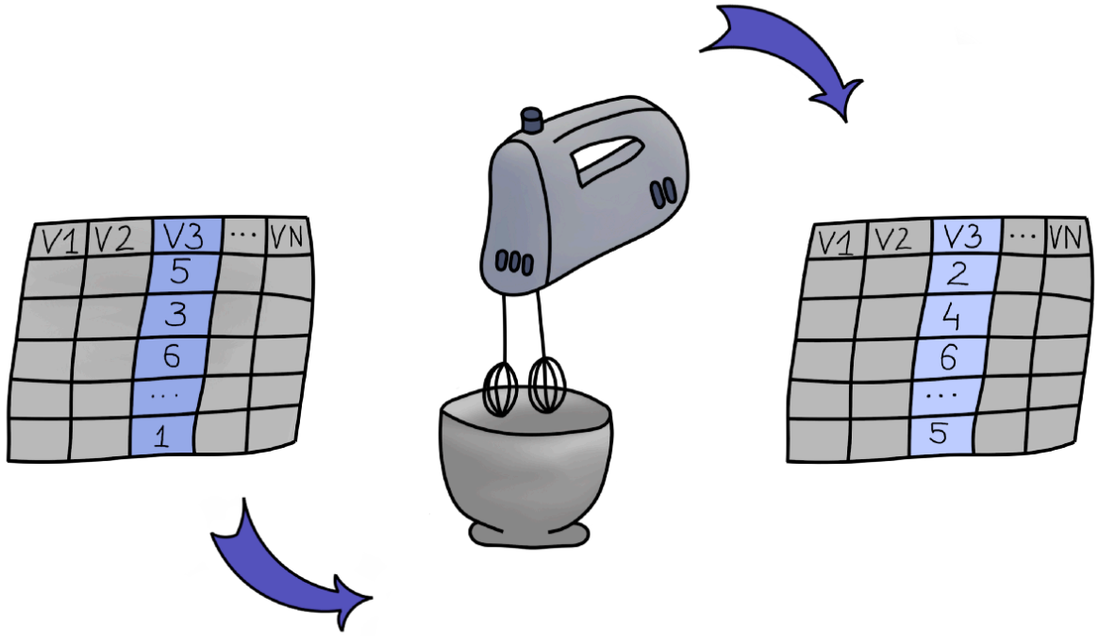
- **Domain-knowledge-based model validation:** identification of the most important variables may be helpful in assessing the validity of the model based on domain knowledge.

- **Knowledge generation:** identification of the most important variables may lead to the discovery of new factors involved in a particular mechanism.

Method

The main idea behind the method (Fisher et al., 2019) is to measure how much does a model's performance change if the effect of a selected explanatory variable, or of a group of variables, is removed?

To remove the effect, we use perturbations, like resampling from an empirical distribution or permutation of the values of the variable.



After permuting the values of the variable, the model's performance will worsen. The larger the change in the performance, the more important is the variable.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2019. "All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *Journal of Machine Learning Research* 20 (177): 1–81. http://jmlr.org/papers/v20/18-760.html.

Algorithm

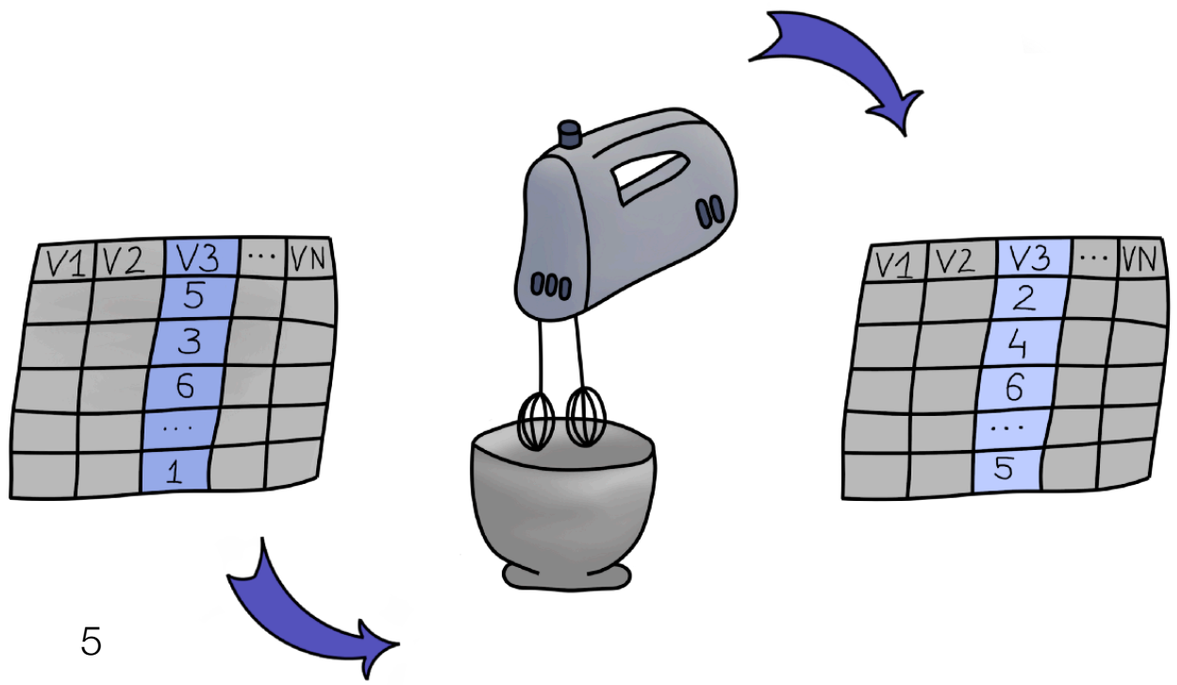
1. Compute $L^0 = L(\hat{y}, X, y)$ the value of the loss function for the original data. Then, for each explanatory variable X^j included in the model, do steps 2-5.

2. Create matrix X^{*j} by permuting the j-th column of X, by permuting the vector of observed values of X^j.

3. Compute the model predictions based $\hat{y}^{*j}$ on the modified data X^{*j}.

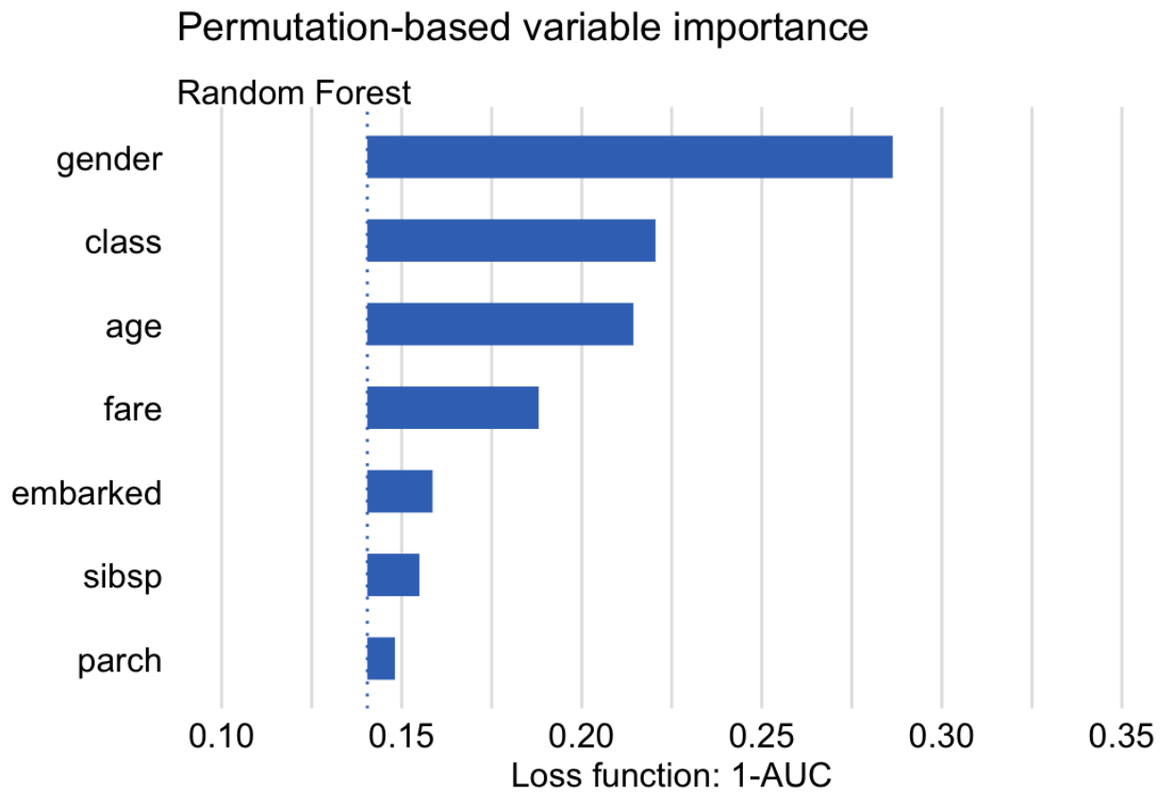
4. Compute the value of the loss function for the modified data $L^{*j} = L(\hat{y}^{*j}, X^{*j}, y)$

5. Quantify the importance of X^j by calculating $vip^j_{diff} = L^{*j} - L^0$ or $vip^j_{diff} = L^{*j}/L^0$.



Single permutation based vip

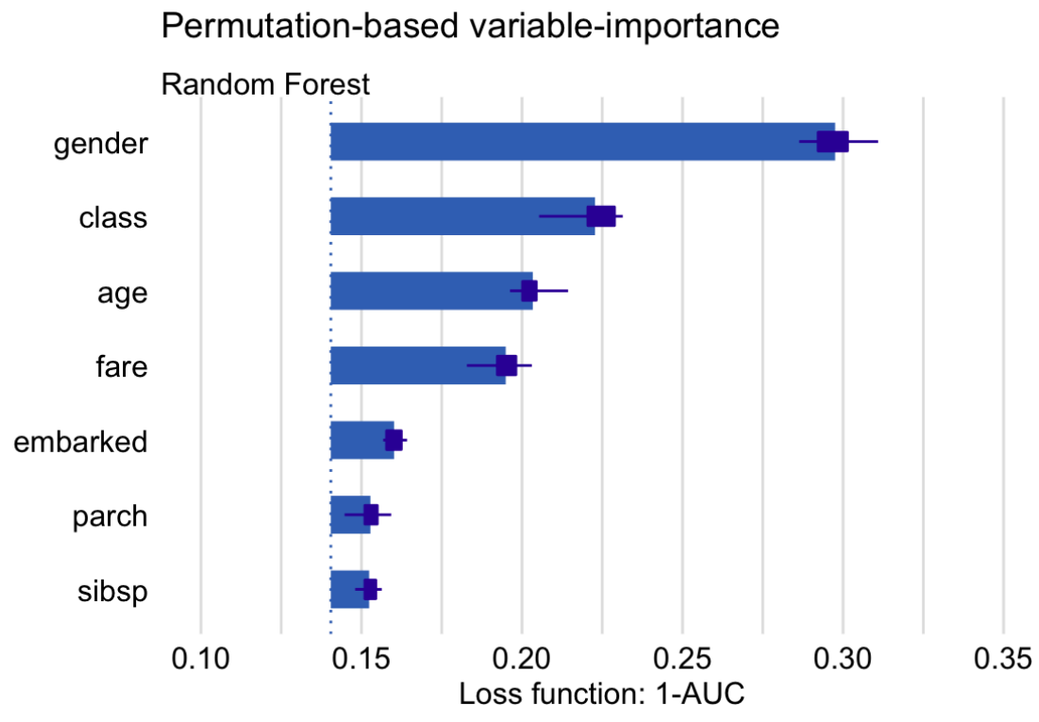
The permutation-based variable-importance evaluation by applying it to the random forest model for the Titanic data:



Vertical dashed-line indicates the L0
The lengths of the bars correspond to vip^{j}_{diff}

k-permutation based vip

To take into account the uncertainty related to the use of permutations, we can consider computing the mean values of over a set of, say, 10 permutations.

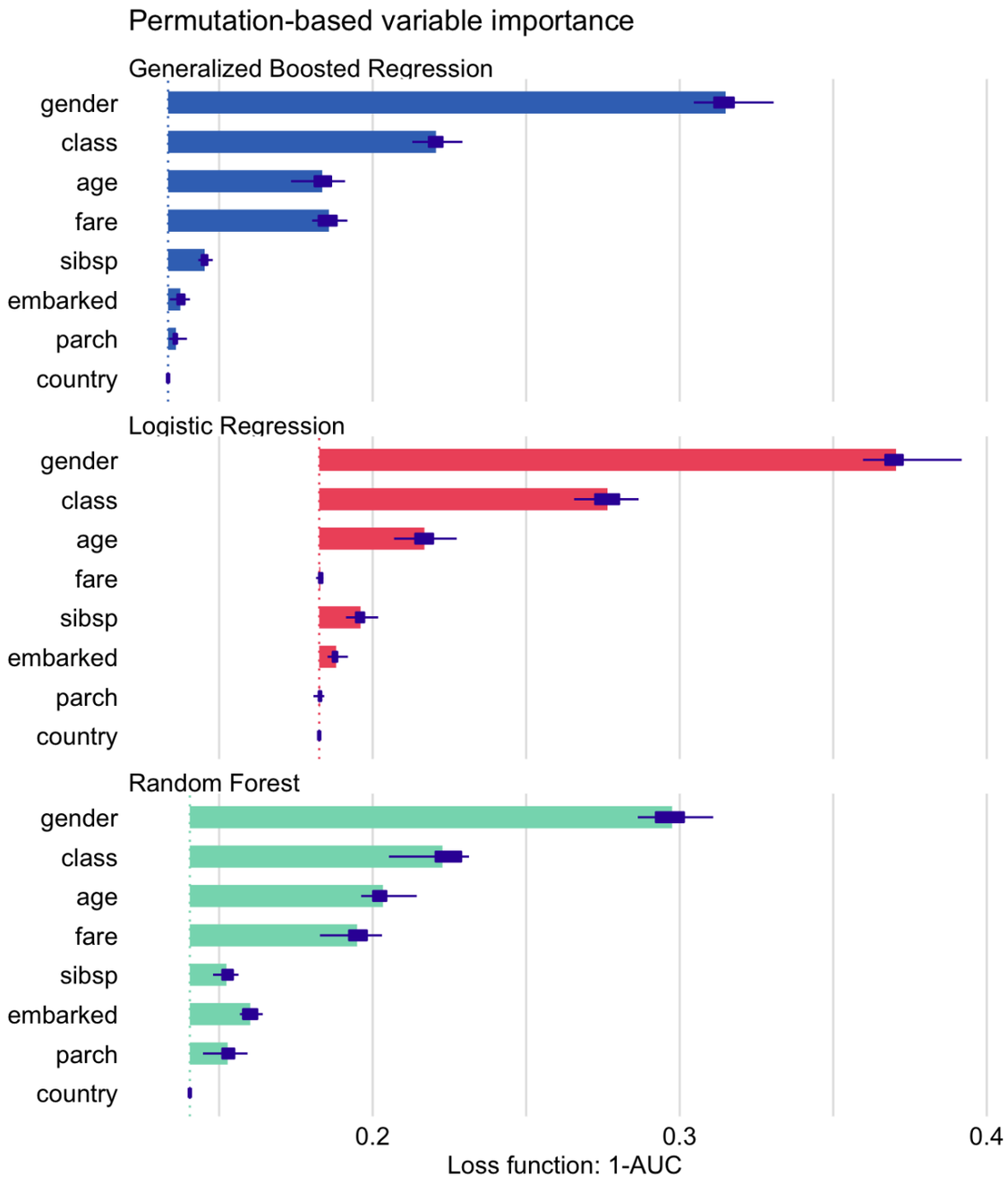


Vertical dashed-line indicates the L0
The lengths of the bars correspond to vip^{j}_{diff}

Comparison of vip in different models

- The importance of the features/variables can be compared in different models for same task.

- The best accurate model, in terms of the smallest value of L0 which are represented by dashed lines, is obtained for generalized boosted regression model.



Pros and Cons

+ model-agnostic
+ plots are easy to understand

- dependence on the random nature of the permutations
- the value of the measure depends on the choice of the loss function

Please feel free to send e-mail about your questions!

📩 mustafa.cavus@pw.edu.pl